Your Project Title (e.g. Replication: Policy Optimization with Demonstrations)

Author names
Department of Computer Science
National Yang Ming Chiao Tung University
{xxx, yyy, zzz}@nycu.edu.tw

1 Problem Overview

This paper introduces a new softmax operator during action selection, which can be used in a SARSA algorithm that computes a Boltzmann policy with a state-dependent temperature parameter. The algorithm is convergent and serves as an alternative to the original Boltzmann policy. The main research problem that is tackled by the paper is that all the common using softmax operator are not idea operator. Some are not non-expansion, some are not differentiable. Hence, The author proposed mellowmax softmax operator that is proved to be an idea operator and generate a well trade-off between exploration and exploitation during action selection.

2 Background and The Algorithm

An ideal softmax operator is a parameterized set of operators that:

- has parameter settings that allow it to approximate maximization arbitrarily accurately to perform reward-seeking behavior;
- is a non-expansion for all parameter settings ensuring convergence to a unique fixed point;
- is differentiable to make it possible to improve via gradient-based optimization; and
- avoids the starvation of non-maximizing actions.
- Let $X = x_1, ..., x_n$ be a vector of values. We define the following operators:

$$\max(\mathbf{X}) = \max_{i \in \{1,\ldots,n\}} x_i \ ,$$

$$\text{mean}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} x_i \ ,$$

$$\text{eps}_\epsilon(\mathbf{X}) = \epsilon \ \text{mean}(\mathbf{X}) + (1 - \epsilon) \max(\mathbf{X}) \ ,$$

$$\text{boltz}_\beta(\mathbf{X}) = \frac{\sum_{i=1}^{n} x_i \ e^{\beta x_i}}{\sum_{i=1}^{n} e^{\beta x_i}} \ .$$

The Boltzmann operator $boltz_\beta(X)$ is differentiable. It also approximates max as $\beta \to \infty$and mean as $\beta \to 0$. However, it is not a non-expansion operator, and therefore, the lack of the non-expansion property leads to multiple fixed points and ultimately a misbehavior in learning and planning.

MellowMax

Author presents a new softmax operator that is similar to the Boltzmann operator yet is a non-expansion operator. Author also proves several critical properties of this new operator, introduce a new softmax policy, and present empirical results. The alternative mellowmax softmax operator is defined as follows:

$$\text{mm}_\omega(\mathbf{X}) = \frac{\log(\frac{1}{n}\sum_{i=1}^{n} e^{\omega x_i})}{\omega} ,$$

Author proves that $mm\omega$ is a non-expansion operator, and therefore, GVI and SARSA under $mm\omega$ are guaranteed to converge to a unique fixed point. Furthermore, the operator acts more and more like pure maximization as the value of ω is increased. Conversely, as ω goes to $-\infty$, the operator approaches the minimum. And as ω gets closer to zero, $mm_\omega(x)$ approaches the mean of the values in X.

Mellowmax Policy

Author formally defines the maximum entropy mellowmax policy of a state s as:

$$\pi_{\text{mm}}(s) = \underset{\pi}{\text{argmin}} \sum_{a\in\mathcal{A}} \pi(a|s)\log\big(\pi(a|s)\big) \qquad (2)$$
$$\text{subject to} \begin{cases} \sum_{a\in\mathcal{A}} \pi(a|s)\hat{Q}(s,a) = \text{mm}_\omega(\hat{Q}(s,.)) \\ \pi(a|s) \geq 0 \\ \sum_{a\in\mathcal{A}} \pi(a|s) = 1 . \end{cases}$$

After using the method of Lagrange multipliers to solve this system of equations, the probability of taking an action under the maximum entropy mellowmax policy has the form:

$$\pi_{mm}(a|s) = \frac{e^{\beta\hat{Q}(s,a)}}{\sum_{a\in\mathcal{A}} e^{\beta\hat{Q}(s,a)}} \quad \forall a \in \mathcal{A} ,$$

where β is a value for which:

$$\sum_{a\in\mathcal{A}} e^{\beta\left(\hat{Q}(s,a) - \text{mm}_\omega\hat{Q}(s,.)\right)}\left(\hat{Q}(s,a) - \text{mm}_\omega\hat{Q}(s,.)\right) = 0 .$$

The argument for the existence of a unique root is simple. As $\beta \to \infty$, the term corresponding to the best action dominates, and so, the function is positive. Conversely, as $\beta \to -\infty$, the term corresponding to the action with lowest utility dominates, and so the function is negative. Finally, by taking the derivative, it is clear that the function is monotonically increasing, allowing us to conclude that there exists only a single root. Therefore, we can find β easily using any root-finding algorithm.

This policy has the same form as Boltzmann softmax, but with a parameter β whose value depends indirectly on ω. This mathematical form arose not from the structure of $mm\omega$, but from maximizing the entropy. One way to view the use of the mellowmax operator, then, is as a form of Boltzmann policy with a temperature parameter chosen adaptively in each state to ensure that the non-expansion property holds.

3 Detailed Implementation

Please explain your implementation in detail. You may do this with the help of pseudo code or a figure of system architecture. Please also highlight which parts of the algorithm lead to the most difficulty in your implementation.

4 Empirical Evaluation

Please showcase your empirical results in this section. Please clearly specify which sets of experiments of the original paper are considered in your report. Please also report the corresponding hyperparameters of each experiment.

5 Conclusion

The whole survey and replication through the paper could summarize the advantages of mellowmax in the following points:

- The non-expansion property: the convergence of GVI is guaranteed
- more stable during the training

and the disadvantages of mellowmax is that:

- Using mellowmax softmax operator may need more time to solve beta, but may get better updates during training.

This paper proposed the mellowmax operator as an alternative to the Boltzmann softmax operator. We also replicated that mellowmax has several desirable properties and that it works favorably in practice. Arguably, mellowmax could be used in place of Boltzmann throughout reinforcement-learning research.

References